

Discussion on the paper: Hypotheses testing by convex optimization by Goldenshluger, Juditsky and Nemirovski

Arnak S. Dalalyan,

3, Avenue Pierre Larousse,
92240 Malakoff, France
e-mail: arnak.dalalyan@ensae.fr
url: <http://arnak-dalalyan.fr/>

This is an exciting piece of work. I agree with the authors that developing computationally tractable methods for hypotheses testing is an important problem in statistics that have received little attention to date. In this discussion, I would like to put the emphasis on three points presented in the paper under discussion that are of particular interest.

Connection with the statistical learning theory

The idea of convexification of the loss function in order to construct computationally tractable procedures has been widely used in statistical learning theory [Zhang, 2004]. In this part of the discussion, I would like to share some thoughts about the similarities of the two approaches.

To this end, let me briefly recall the principle of loss convexification in the problem of binary classification. One observes n iid pairs $\{(X_i, Y_i)\}_{i=1, \dots, n}$ drawn from an unknown distribution P on the product space $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{-1, +1\}$ and the goal is to design a prediction rule $g : \mathcal{X} \rightarrow \mathcal{Y}$ with the smallest possible misclassification error rate

$$R_P(g) = \mathbf{E}_P[\mathbf{1}(Y \neq g(X))] = \mathbf{E}_P[\mathbf{1}(-Yg(X) \geq 0)]. \quad (1)$$

The convexification is achieved in two steps. First, the classification risk is replaced by the ϕ -risk

$$A_P(g) = \mathbf{E}_P[\phi(-Yg(X))], \quad (2)$$

where $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is a convex function often referred to as the convex surrogate loss. Second, the set of “pure” classification rules $g : \mathcal{X} \rightarrow \mathcal{Y}$ is extended to “generalized” rules $h : \mathcal{X} \rightarrow \mathbb{R}$ with the convention that the predictions furnished by h and $\text{sgn}(h)$ are the same. The ϕ -risk is accordingly extended to all generalized prediction rules: $A_P(h) = \mathbf{E}_P[\phi(-Yh(X))]$. As a consequence of this construction, if \mathcal{H} is a convex subset of the set of all measurable functions from \mathcal{X} to \mathbb{R} , then the computation of the empirical risk minimizer (ERM)

$$\hat{h}_{n, \mathcal{H}} \in \arg \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \phi(-Y_i h(X_i)) \quad (3)$$

amounts to solving a convex program. The most common choices for the function ϕ are the hinge loss $\phi(u) = (1 + u)_+$, the exponential loss $\phi(u) = e^u$ and the logistic loss $\phi(u) = \log(1 + e^u)$.

Let me turn now to the problem of testing two hypotheses Θ_0 and Θ_1 based on n observations X_1, \dots, X_n independently drawn from a distribution P_{θ^*} on \mathcal{X} . Let $\Theta = \Theta_0 \cup \Theta_1$ and $s : \Theta \rightarrow \{\pm 1\}$ be the function that equals -1 on the set Θ_0 and $+1$ on the set Θ_1 . The usual loss of a pure test $T : \mathcal{X}^n \rightarrow \{\pm 1\}$ associated with a sample $\mathbf{X}^{(n)} = (X_1, \dots, X_n)$ drawn from $P_{\theta^*}^n := P_{\theta^*}^{\otimes n}$ is

$$\ell(T(\mathbf{X}^{(n)}), \theta^*) = \mathbb{1}(T(\mathbf{X}^{(n)}) \neq s(\theta^*)) = \mathbb{1}(-T(\mathbf{X}^{(n)})s(\theta^*) \geq 0).$$

The corresponding risk is $R_{P_{\theta^*}}(T) = \mathbf{E}_{\theta^*}[\ell(T(\mathbf{X}^{(n)}), \theta^*)]$ and the worst case risk is

$$\begin{aligned} \epsilon(T) &= \sup_{\theta \in \Theta} R_{P_{\theta^*}}(T) = \sup_{\theta \in \Theta} P_{\theta}(T(\mathbf{X}^{(n)}) \neq s(\theta)) \\ &= \sup_{\theta \in \Theta} \mathbf{E}_{\theta}[\mathbb{1}(-T(\mathbf{X}^{(n)})s(\theta) \geq 0)]. \end{aligned} \quad (4)$$

Comparing (4) with (1), one can see some clear similarities between the problems of finding binary predictors g minimizing the misclassification error rate and that of finding testing procedures T minimizing the worst case error rate $\epsilon(T)$. In both problems the decision rules form a nonconvex set and the performance measure is defined as the expected loss for a nonconvex loss function (the Heaviside step function). However, there is an important difference consisting in the fact that—contrary to (1)—the expectation at the right-hand side of (4) does not admit an empirical counterpart that is easily computable from the sample. Therefore, even if one applies the aforementioned two steps of convexification, this does not readily yield a test procedure computable by solving a convex program (in the spirit of (3)).

Elaborating on these ideas, one can define the following convexified strategy for testing the hypothesis Θ_0 against Θ_1 . Given a convex subset \mathcal{H} of the set of measurable functions from \mathcal{X}^n to \mathbb{R} and a convex loss $\phi : \mathbb{R} \rightarrow \mathbb{R}$, define

$$\hat{h}_{n, \mathcal{H}}^{\phi} \in \arg \min_{h \in \mathcal{H}} \sup_{\theta \in \Theta} G_{\phi}(h, \theta), \quad G_{\phi}(h, \theta) = \mathbf{E}_{\theta}[\phi(-h(\mathbf{X}^{(n)})s(\theta))]. \quad (5)$$

In this “saddle-point” formulation, the outer minimisation problem has the attractive property of being convex: it has a convex feasible set and a convex cost function. Unfortunately, in general, the inner maximization problem is not concave and there is no particular reason to expect that it can be efficiently solved for any given h when the dimensionality of θ is large. To circumvent this drawback, the authors had the ingenious idea to combine the following three facts:

- the saddle point of $G(h, \theta)$ coincides with the saddle point of $\log G(h, \theta)$,
- when the model $\{P_{\theta} : \theta \in \Theta\}$ belongs to an exponential family, it is natural to choose \mathcal{H} as the linear span of the sufficient statistics: $\mathcal{H}_0 = \text{Span}(S_j : j = 1, \dots, m)$,
- for some statistical models¹ belonging to an exponential family, for every $h \in \mathcal{H}_0$, the mapping $\theta \mapsto \log(\mathbf{E}_{\theta}[\exp(-h(\mathbf{X}^{(n)})s(\theta))])$ is concave.

¹It could be helpful to mention that the concavity property holds for the usual parameterization and does not hold for the parameterization in terms of the natural parameters in the sense of exponential families.

This leads to the test procedure

$$\hat{h}_{n, \mathcal{H}_0}^{\text{exp}} \in \arg \min_{h \in \mathcal{H}_0} \sup_{\theta \in \Theta} \log G_{\text{exp}}(h, \theta), \quad G_{\text{exp}}(h, \theta) = \mathbf{E}_{\theta}[e^{-h(\mathbf{X}^{(n)})s(\theta)}]. \quad (6)$$

The final step of construction aims at convexifying the feasible set of the inner maximization problem. In the case when $\Theta = \Theta_0 \cup \Theta_1$ with convex sets Θ_0 and Θ_1 , this aim is achieved by replacing $\sup_{\theta \in \Theta} \log G_{\text{exp}}(h, \theta)$ by the expression $\sup_{(\theta, \bar{\theta}) \in \Theta_0 \times \Theta_1} \log G_{\text{exp}}(h, \theta) + \log G_{\text{exp}}(h, \bar{\theta})$, which does not impact the error of testing too much in view of the inequalities

$$\begin{aligned} \sup_{\theta \in \Theta} \log G_{\text{exp}}(h, \theta) &\leq \sup_{(\theta, \bar{\theta}) \in \Theta_0 \times \Theta_1} \{ \log G_{\text{exp}}(h, \theta) + \log G_{\text{exp}}(h, \bar{\theta}) \} \\ &\leq 2 \sup_{\theta \in \Theta} \log G_{\text{exp}}(h, \theta). \end{aligned}$$

An important remark to be made here is that—in the case of exponential loss ϕ —taking the logarithm of G_{ϕ} does not break the convexity with respect to h . So, in this notation, the test proposed and studied by the authors is

$$\tilde{h}_{n, \mathcal{H}_0}^{\text{exp}} \in \arg \min_{h \in \mathcal{H}_0} \sup_{(\theta, \bar{\theta}) \in \Theta_0 \times \Theta_1} \{ \log G_{\text{exp}}(h, \theta) + \log G_{\text{exp}}(h, \bar{\theta}) \}. \quad (7)$$

I believe that these explanations shed some additional light on the construction proposed in Theorem 2.1 of the paper under discussion. This also raises several questions that might be interesting to investigate in the future. In particular, a compelling question is to characterize the set of surrogate loss functions ϕ that lead to computationally tractable testing procedures and for which the testing error rate remains small. Another question is the possibility to deal with test (6) directly, without using the final step of convexification. At a heuristic level, the risk of $\hat{h}_{n, \mathcal{H}_0}^{\text{exp}}$ should be smaller than that of $\tilde{h}_{n, \mathcal{H}_0}^{\text{exp}}$. Therefore, the advantage of the latter would be only computational tractability. I wonder if it is possible to efficiently compute the test $\hat{h}_{n, \mathcal{H}_0}^{\text{exp}}$, despite the lack of convex-concavity of the cost function, exploiting the facts that (a) for every h , the sup of $\log G_{\text{exp}}(h, \theta)$ over Θ can be efficiently computed, and (b) for every θ , the minimum of $\log G_{\text{exp}}(h, \theta)$ over \mathcal{H}_0 can be efficiently computed as well.

Reduction to testing simple hypotheses

The definition of the test given by the authors in Theorem 2.1, see also Eq. (7) above, is well suited for the computational purposes but, in my opinion, has the inconvenience of hiding the main reason why the proposed test is a natural one to use in the setting under consideration. In fact, the proposed test can be alternatively defined as follows: in order to distinguish between two (convex) hypotheses Θ_0 and Θ_1 based on a sample $\mathbf{X} \sim P_{\theta^*}$,

1. Determine the two closest points $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$ in terms of the Hellinger distance between the corresponding distributions (in other terms, find the two representers P_{θ_0} and P_{θ_1} in the families $\{P_{\theta} : \theta \in \Theta_0\}$ and $\{P_{\theta} : \theta \in \Theta_1\}$ that are the hardest to distinguish). This step is completely data independent.
2. Apply the standard likelihood-ratio test to the problem of choosing among two simple hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.

The equivalence of these two definitions follows from the proof of Theorem 2.1, see Eq. (52). In Section 2.3.2, this interpretation is presented for the discrete observation scheme. At a conceptual level, it is important to underline that the same interpretation holds true in the general case as well. However, from a practical point of view, the definition given in the paper is more convenient than the foregoing one since the first step of the latter, generally, is not computationally tractable.

Testing error for inexact solutions

As it is judiciously noted by the authors, in many practical situations, the exact computation of the saddle point in (7) can not be performed. Then, one relies on an approximation of the saddle point and it is a central task to assess how this approximation error impacts the error of testing. I find it relevant to measure the error of approximation in terms of the magnitude of violation of first-order optimality conditions (see, for instance, Eq. (8) of the paper under discussion). In such a context, the authors establish upper bounds on the error of the test based on an approximate solution to the saddle point problem. For example, in the case of the Gaussian observation scheme explored in Section 2.3.1, it is shown that the worst-case error rate of the test based on the exact solution is

$$\varepsilon_* = 1 - \Phi\left(\frac{1}{2}\|\Sigma^{-1/2}(\theta_0 - \theta_1)\|_2\right), \quad (8)$$

where Φ is the cumulative distribution function of the standard normal distribution and (θ_0, θ_1) is the second argument of the solution to the saddle point problem. On the other hand, when an inexact solution $(\tilde{\theta}_0, \tilde{\theta}_1)$ is used, with an approximation error bounded by $\delta > 0$, the worst-case error rate satisfies (see Eq. (9)):

$$\tilde{\varepsilon} \leq 1 - \Phi\left(\frac{1}{2}\|\Sigma^{-1/2}(\tilde{\theta}_0 - \tilde{\theta}_1)\|_2 - \frac{\delta}{\|\Sigma^{-1/2}(\tilde{\theta}_0 - \tilde{\theta}_1)\|_2}\right).$$

In my opinion, it is worth complementing this upper bound by another one that involves only the exact solution (θ_0, θ_1) and, therefore, makes it easier to compare the two errors ε_* and $\tilde{\varepsilon}$. In the case of Gaussian observation scheme, this can be easily done. In fact, one can deduce from the first-order exact and approximate optimality conditions that

$$\|\Sigma^{-1/2}(\theta_0 - \theta_1)\|_2 - \sqrt{\delta} \leq \|\Sigma^{-1/2}(\tilde{\theta}_0 - \tilde{\theta}_1)\|_2 \leq \|\Sigma^{-1/2}(\theta_0 - \theta_1)\|_2 + \sqrt{\delta} \quad (9)$$

Since the Gaussian cdf is increasing, we infer from this inequality that

$$\tilde{\varepsilon} \leq 1 - \Phi\left(\frac{1}{2}\|\Sigma^{-1/2}(\theta_0 - \theta_1)\|_2 - \frac{\sqrt{\delta}}{2} - \frac{\delta}{\|\Sigma^{-1/2}(\theta_0 - \theta_1)\|_2 + \sqrt{\delta}}\right).$$

An even more elegant bound can be obtained if the normalized approximate optimality condition is used: $\forall(\theta, \bar{\theta}) \in \Theta_0 \times \Theta_1$, it holds

$$(\tilde{\theta}_1 - \tilde{\theta}_0)\Sigma^{-1}(\theta - \tilde{\theta}_0) + (\tilde{\theta}_0 - \tilde{\theta}_1)\Sigma^{-1}(\bar{\theta} - \tilde{\theta}_1) \leq \delta\|\Sigma^{-1/2}(\tilde{\theta}_0 - \tilde{\theta}_1)\|_2^2.$$

In this case, inequalities (9) take the form

$$\frac{\|\Sigma^{-1/2}(\theta_0 - \theta_1)\|_2}{1 + \sqrt{\delta}} \leq \|\Sigma^{-1/2}(\tilde{\theta}_0 - \tilde{\theta}_1)\|_2 \leq \frac{\|\Sigma^{-1/2}(\theta_0 - \theta_1)\|_2}{1 - \sqrt{\delta}} \quad (10)$$

and we get

$$\tilde{\varepsilon} \leq 1 - \Phi \left\{ \left(\frac{1}{2} - \delta \right) \frac{\|\Sigma^{-1/2}(\theta_0 - \theta_1)\|_2}{1 + \sqrt{\delta}} \right\}.$$

This inequality allows for an easy comparison of $\tilde{\varepsilon}$ and ϵ_* in the case of Gaussian observations. In the case of other observation schemes, deriving this type of upper bounds seems to be more challenging and constitutes an interesting avenue of future research.

Acknowledgments

The research of the author is partially supported by the grant Investissements d’Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

References

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.